$L_1$ Penalization
The Lasso and Adaptive Lasso

Justin Petrovich$^1$

$^1$Department of Statistics

October 7, 2016
Outline

1 Lasso
   - Background & Motivation
   - Definition & Properties
   - Examples
   - Advantages & Disadvantages

2 Adaptive Lasso
   - Problem: Shortcoming of Lasso
   - Definition of Adaptive Lasso
   - Properties
   - Comparisons to Other Procedures
Motivation

- Prostate Cancer Data: PSA level used as a rough threshold for determining patients at risk of prostate cancer. Data on 8 patient factors to try to predict PSA level.
- Goals in regression: (1) Prediction accuracy and (2) model interpretability.
- Model options?
Motivation

- OLS regression?
- Ridge regression?
- Best subset regression?
- Would like to find a method to perform variable selection while maintaining good predictive accuracy.
Brief History

- Breiman (1993): Non-negative garotte
- Frank and Friedman (1993): Bridge Regression
- Tibshirani (1996): Lasso Regression
Brief History (continued)

\[
\hat{\beta}_{\text{ridge}, \lambda} = \arg\min_{\beta} \sum_{i=1}^{N} \left( y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \quad (1)
\]

\[
\hat{\beta}_{\text{nng}}^j = c_j \hat{\beta}_{\text{ols}}^j \quad \text{such that} \quad (2)
\]

\[
c = \arg\min_{c_j \geq 0} \sum_{i=1}^{N} \left( y_i - \sum_{j=1}^{p} c_j \hat{\beta}_{\text{ols}}^j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} c_j
\]

\[
\hat{\beta}_{\text{bridge}, \lambda} = \arg\min_{\beta} \sum_{i=1}^{N} \left( y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^\gamma \quad (3)
\]
Definition: Lasso

Given data of the form \((x^i, y_i), i = 1, 2, \cdots, N\), where \(x^i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T\) are the \(p\) predictor variables and \(y_i\) are the responses, assuming independent observations, and assuming scaled and centered data where \(\sum_i x_{ij}/N = 0\) and \(\sum_i x_{ij}^2/N = 1\), we define the Lasso coefficient estimates of \(\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)^T\) and \(\hat{\alpha}\) to be

\[
(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p} \left[ \sum_{i=1}^{N} \left( y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right] \quad \text{subject to } \sum_j |\beta_j| \leq t
\]

where \(\alpha\) is the intercept of the model.
Definition: Lasso

Ignoring the intercept (which we can do WLOG) and writing in the Lagrangian form, we can rephrase this to be more similar to the first 3 equations:

\[ \hat{\beta}_{\text{lasso}} = \arg\min_\beta \sum_{i=1}^N \left( y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \]  \hspace{1cm} (4)

Tuning parameter: \( \lambda \) (or, equivalently \( t \) from the previous definition).
Solution to Lasso utilizes soft thresholding:

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^{ols})(|\hat{\beta}_j^{ols}| - \lambda)^+$$

where \((x)^+ = x I_{\{x>0\}}\) and \(\lambda\) is chosen by the condition \(\sum |\hat{\beta}| = t\).
Bias-Variance Trade-off

**Figure:** Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on a simulated data set. Figure can be found on page 223 of *An Introduction to Statistical Learning* by James et. al.
Lasso vs. Ridge Regression

- Both provide variable shrinkage.
- Both can improve variance over OLS estimates, which would improve prediction accuracy overall.
- Lasso better suited for sparse settings (small number of non-zero coefficients).
- Ridge better suited for situations when there are many predictors of similar size.
- Neither Ridge nor Lasso universally dominate the other.
### Table 1

**Results for the prostate cancer example**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Least squares results</th>
<th></th>
<th>Subset selection results</th>
<th></th>
<th>Lasso results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>Standard error</td>
<td>Z-score</td>
<td>Coefficient</td>
<td>Standard error</td>
</tr>
<tr>
<td>1 intcpt</td>
<td>2.48</td>
<td>0.07</td>
<td>34.46</td>
<td>2.48</td>
<td>0.07</td>
</tr>
<tr>
<td>2 lcavol</td>
<td>0.69</td>
<td>0.10</td>
<td>6.68</td>
<td>0.65</td>
<td>0.09</td>
</tr>
<tr>
<td>3 lweight</td>
<td>0.23</td>
<td>0.08</td>
<td>2.67</td>
<td>0.25</td>
<td>0.07</td>
</tr>
<tr>
<td>4 age</td>
<td>-0.15</td>
<td>0.08</td>
<td>-1.76</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>5 lbph</td>
<td>0.16</td>
<td>0.08</td>
<td>1.83</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>6 svi</td>
<td>0.32</td>
<td>0.10</td>
<td>3.14</td>
<td>0.28</td>
<td>0.09</td>
</tr>
<tr>
<td>7 lcp</td>
<td>-0.15</td>
<td>0.13</td>
<td>-1.16</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>8 gleason</td>
<td>0.03</td>
<td>0.11</td>
<td>0.29</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>9 pgg45</td>
<td>0.13</td>
<td>0.12</td>
<td>1.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

---

**Figure:** page 274 of *Regression Shrinkage and Selection via the Lasso* by Tibshirani
Simulations
Example 1

- Simulated 50 data sets, each with $n = 20$, from

$$y = \beta^T x + \sigma \epsilon \quad \epsilon \sim N(0, 1).$$

- $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$

- $\sigma = 3$ and $\text{cor}(x_i, x_j) = 0.5^{|i-j|}$. 
### Example 1

**TABLE 3**

*Results for example 1†*

<table>
<thead>
<tr>
<th>Method</th>
<th>Median mean-squared error</th>
<th>Average no. of 0 coefficients</th>
<th>Average $\hat{s}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least squares</td>
<td>2.79 (0.12)</td>
<td>0.0</td>
<td>—</td>
</tr>
<tr>
<td>Lasso (cross-validation)</td>
<td>2.43 (0.14)</td>
<td>3.3</td>
<td>0.63 (0.01)</td>
</tr>
<tr>
<td>Lasso (Stein)</td>
<td>2.07 (0.10)</td>
<td>2.6</td>
<td>0.69 (0.02)</td>
</tr>
<tr>
<td>Lasso (generalized cross-validation)</td>
<td>1.93 (0.09)</td>
<td>2.4</td>
<td>0.73 (0.01)</td>
</tr>
<tr>
<td>Garotte</td>
<td>2.29 (0.16)</td>
<td>3.9</td>
<td>—</td>
</tr>
<tr>
<td>Best subset selection</td>
<td>2.44 (0.16)</td>
<td>4.8</td>
<td>—</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>3.21 (0.12)</td>
<td>0.0</td>
<td>—</td>
</tr>
</tbody>
</table>

†Standard errors are given in parentheses.

**Figure:** page 279 of *Regression Shrinkage and Selection via the Lasso* by Tibshirani
Simulated 50 data sets, each with $n = 20$, from

$$y = \beta^T x + \sigma \epsilon \quad \epsilon \sim N(0, 1).$$

- $\beta = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)^T$
- $\sigma = 3$ and $\text{cor}(x_i, x_j) = 0.5|i-j|$. 
Example 2

TABLE 6
Results for example 2†

<table>
<thead>
<tr>
<th>Method</th>
<th>Median mean-squared error</th>
<th>Average no. of 0 coefficients</th>
<th>Average $$\hat{s}$$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least squares</td>
<td>6.50 (0.64)</td>
<td>0.0</td>
<td>—</td>
</tr>
<tr>
<td>Lasso (cross-validation)</td>
<td>5.30 (0.45)</td>
<td>3.0</td>
<td>0.50 (0.03)</td>
</tr>
<tr>
<td>Lasso (Stein)</td>
<td>5.85 (0.36)</td>
<td>2.7</td>
<td>0.55 (0.03)</td>
</tr>
<tr>
<td>Lasso (generalized cross-validation)</td>
<td>4.87 (0.35)</td>
<td>2.3</td>
<td>0.69 (0.23)</td>
</tr>
<tr>
<td>Garotte</td>
<td>7.40 (0.48)</td>
<td>4.3</td>
<td>—</td>
</tr>
<tr>
<td>Subset selection</td>
<td>9.05 (0.78)</td>
<td>5.2</td>
<td>—</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>2.30 (0.22)</td>
<td>0.0</td>
<td>—</td>
</tr>
</tbody>
</table>

†Standard errors are given in parentheses.

Figure: page 282 of *Regression Shrinkage and Selection via the Lasso* by Tibshirani
Simulations
Example 3

- Simulated 50 data sets, each with $n = 20$, from

$$y = \beta^T x + \sigma \epsilon \quad \epsilon \sim N(0, 1).$$

- $\beta = (5, 0, 0, 0, 0, 0, 0, 0)^T$
- $\sigma = 2$ and $\text{cor}(x_i, x_j) = 0.5 |i-j|$. 
Example 3

**TABLE 7**

*Results for example 3†*

<table>
<thead>
<tr>
<th>Method</th>
<th>Median mean-squared error</th>
<th>Average no. of 0 coefficients</th>
<th>Average $\hat{s}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least squares</td>
<td>2.89 (0.04)</td>
<td>0.0</td>
<td>—</td>
</tr>
<tr>
<td>Lasso (cross-validation)</td>
<td>0.89 (0.01)</td>
<td>3.0</td>
<td>0.50 (0.03)</td>
</tr>
<tr>
<td>Lasso (Stein)</td>
<td>1.26 (0.02)</td>
<td>2.6</td>
<td>0.70 (0.01)</td>
</tr>
<tr>
<td>Lasso (generalized cross-validation)</td>
<td>1.02 (0.02)</td>
<td>3.9</td>
<td>0.63 (0.04)</td>
</tr>
<tr>
<td>Garotte</td>
<td>0.52 (0.01)</td>
<td>5.5</td>
<td>—</td>
</tr>
<tr>
<td>Subset selection</td>
<td>0.64 (0.02)</td>
<td>6.3</td>
<td>—</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>3.53 (0.05)</td>
<td>0.0</td>
<td>—</td>
</tr>
</tbody>
</table>

†Standard errors are given in parentheses.

**Figure:** page 282 of * Regression Shrinkage and Selection via the Lasso* by Tibshirani
Simulations
Example 4

- Simulated 50 data sets, each with $n = 100$ and $p = 40$, from

$$y = \beta^T x + 15\epsilon \quad \epsilon \sim N(0, 1).$$

- $x_{ij} = z_{ij} + z_i$ where $z_{ij}$ and $z_i$ are independent standard normal variates.

- $\beta = (0, \cdots, 0, 2, \cdots, 2, 0, \cdots, 0, 2, \cdots, 2)^T$

- First 10 true $\beta$ coefficients are 0, next 10 are 2, next 10 are 0, and the final 10 are 0.
**Example 4**

<table>
<thead>
<tr>
<th>Method</th>
<th>Median mean-squared error</th>
<th>Average no. of 0 coefficients</th>
<th>Average $\hat{s}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least squares</td>
<td>137.3 (7.3)</td>
<td>0.0</td>
<td>—</td>
</tr>
<tr>
<td>Lasso (Stein)</td>
<td>80.2 (4.9)</td>
<td>14.4</td>
<td>0.55 (0.02)</td>
</tr>
<tr>
<td>Lasso (generalized cross-validation)</td>
<td>64.9 (2.3)</td>
<td>13.6</td>
<td>0.60 (0.88)</td>
</tr>
<tr>
<td>Garotte</td>
<td>94.8 (3.2)</td>
<td>22.9</td>
<td>—</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>57.4 (1.4)</td>
<td>0.0</td>
<td>—</td>
</tr>
</tbody>
</table>

†Standard errors are given in parentheses.

**Figure:** page 283 of *Regression Shrinkage and Selection via the Lasso* by Tibshirani
We summarize the results from the simulations:

- *small number of large effects:* Subset selection and the garotte are best, lasso beats ridge

- *small to moderate number of moderate-sized effects:* Lasso generally does best with ridge and the garotte close, but all are comparable

- *large number of small effects:* Ridge regression the clear winner, then lasso, garotte, subset
The lasso is competitive with the garotte and Ridge regression in terms of predictive accuracy, and has the added advantage of producing interpretable models by shrinking coefficients to exactly 0. It performs continuous shrinkage, avoiding the drawback of subset selection. It is well-suited for sparse settings, in which only a small number of potential variables should be kept, and the others require at least some shrinkage.
While the Lasso seems like a very viable procedure so far, a very desirable property is lacking. Specifically, we do not know whether the Lasso has the oracle property.
Definition: Oracle Property

Following the definition of Fan and Li (2001), $\delta$ is an oracle procedure if $\hat{\beta}(\delta)$ has the following properties:

- Identifies the right subset model, $\{j : \hat{\beta}_j \neq 0\} = \mathcal{A}$
- Has the optimal estimation rate,
  \[ \sqrt{n}(\hat{\beta}(\delta)_{\mathcal{A}} - \beta^*_\mathcal{A}) \overset{d}{\rightarrow} N(0, \Sigma^*) \], where $\Sigma^*$ is the covariance matrix knowing the true subset model.

**Note:** for data with $p$ true predictors, $\beta_1^*, \beta_2^*, \cdots, \beta_p^*$, we define $\mathcal{A} = \{j : \beta_j^* \neq 0\}$ and assume that $|\mathcal{A}| = p_0 < p$. 
Set-up

- \( y_i = \mathbf{x}_i \beta^* + \epsilon_i \), where \( \epsilon_1, \epsilon_2, \cdots, \epsilon_n \) are iid, mean 0 random variables with variance \( \sigma^2 \).
- \( \frac{1}{n} \mathbf{X}^T \mathbf{X} \rightarrow \mathbf{C} \), where \( \mathbf{C} \) is a positive definite matrix.
- Assume \( \mathcal{A} = \{1, 2, \cdots, p_0\} \).
- Let \( \mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} \) where \( \mathbf{C}_{11} \) is a \( p_0 \times p_0 \) matrix.
- The Lasso estimates are
  \[
  \hat{\beta}^{(n)} = \arg\min_{\beta} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j \right\|_2^2 + \lambda_n \sum_{j=1}^{p} |\beta_j| \]
  where \( \lambda_n \) varies with \( n \). Call \( \mathcal{A}_n = \{j : \hat{\beta}_j^{(n)} \neq 0\} \).
- Lasso variable selection is consistent iff \( \lim_{n} P(\mathcal{A}_n = \mathcal{A}) = 1 \).
Theorem 1 (Necessary Condition). Suppose that $\lim_n P(\mathcal{A}_n = \mathcal{A}) = 1$. Then there exists some sign vector $\mathbf{s} = (s_1, \cdots, s_{p_0})^T$, $s_j = 1$ or $-1$, such that

$$|\mathbf{C}_{21} \mathbf{C}_{11}^{-1} \mathbf{s}| \leq 1.$$  \hspace{1cm} (5)

The above inequality is understood componentwise.
On Consistency of the Lasso Estimates

- If equation (5) is not satisfied, lasso procedure is inconsistent.

- Furthermore, condition (5) is nontrivial: Zou (2006) provides an example of when the condition fails, in Corollary 1.

- Consistency of lasso is guaranteed under certain conditions (e.g. orthogonal design, or given proper choice of $\lambda_n$ when $p = 2$).
Idea: introduce weights to the penalty on each coefficient in the lasso procedure. Call this new procedure the *Adaptive Lasso*:

\[
\hat{\beta}^{*}(n) = \arg\min_{\beta} \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j |\beta_j| \tag{6}
\]

and let \( A^*_n = \{ j : \hat{\beta}^{*}(n)_j \neq 0 \} \). We define the weight vector as \( \hat{w} = 1/|\hat{\beta}|^\gamma \), where \( \hat{\beta} \) is a root-n-consistent estimator to \( \beta^* \), such as \( \hat{\beta}(\text{ols}) \), and \( \gamma > 0 \).
Theorem 2 (Oracle properties). Suppose that $\lambda_n / \sqrt{n} \to 0$ and $\lambda_n n(\gamma^{-1})/2 \to \infty$. Then the adaptive lasso estimates must satisfy the following:

- Consistency in variable selection: $\lim_n P(A^*_n = \mathcal{A}) = 1$
- Asymptotic normality: $\sqrt{n}(\hat{\beta}^*_\mathcal{A} - \beta^*_\mathcal{A}) \xrightarrow{d} N(0, \sigma^2 \times \mathbf{C}_{11}^{-1})$. 
Simulation Set-up

Model 1: \( \beta = (3, 1.5, 0, 0, 2, 0, 0, 0) \); \( \mathbf{x}_i \) are iid normal vectors for \( i = 1, 2, \cdots, n \); pairwise correlation between \( \mathbf{x}_{j_1} \) and \( \mathbf{x}_{j_2} \) is \( \text{cor}(j_1, j_2) = 0.5|j_1 - j_2| \); \( \sigma = 1, 3, 6 \) and \( n = 20, 60 \).

Model 2: Same as model 1, but \( \beta_j = 0.85 \) for all \( j \). Again, \( \sigma = 1, 3, 6 \), but now \( n = 40, 80 \).
## Simulation Comparison

Table 2. Simulation Models 1 and 2, Comparing the Median RPE Based on 100 Replications

<table>
<thead>
<tr>
<th></th>
<th>Model 1 (n = 20)</th>
<th></th>
<th></th>
<th>Model 1 (n = 60)</th>
<th></th>
<th></th>
<th>Model 2 (n = 40)</th>
<th></th>
<th></th>
<th>Model 2 (n = 80)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma = 1$</td>
<td>$\sigma = 3$</td>
<td>$\sigma = 6$</td>
<td>$\sigma = 1$</td>
<td>$\sigma = 3$</td>
<td>$\sigma = 6$</td>
<td>$\sigma = 1$</td>
<td>$\sigma = 3$</td>
<td>$\sigma = 6$</td>
<td>$\sigma = 1$</td>
</tr>
<tr>
<td>Lasso</td>
<td>.414(.046)</td>
<td>.395(.039)</td>
<td>.275(.026)</td>
<td>.103(.008)</td>
<td>.102(.008)</td>
<td>.107(.012)</td>
<td>.205(.015)</td>
<td>.214(.014)</td>
<td>.161(.009)</td>
<td>.094(.008)</td>
</tr>
<tr>
<td>Adaptive lasso</td>
<td>.261(.023)</td>
<td>.369(.029)</td>
<td>.336(.031)</td>
<td>.073(.004)</td>
<td>.094(.012)</td>
<td>.117(.008)</td>
<td>.203(.015)</td>
<td>.237(.016)</td>
<td>.190(.008)</td>
<td>.093(.007)</td>
</tr>
<tr>
<td>SCAD</td>
<td>.218(.029)</td>
<td>.508(.044)</td>
<td>.428(.019)</td>
<td>.053(.008)</td>
<td>.104(.016)</td>
<td>.119(.014)</td>
<td>.223(.018)</td>
<td>.297(.028)</td>
<td>.230(.009)</td>
<td>.096(.007)</td>
</tr>
<tr>
<td>Garotte</td>
<td>.227(.007)</td>
<td>.488(.043)</td>
<td>.385(.030)</td>
<td>.069(.006)</td>
<td>.102(.008)</td>
<td>.118(.009)</td>
<td>.199(.018)</td>
<td>.273(.024)</td>
<td>.219(.019)</td>
<td>.095(.006)</td>
</tr>
</tbody>
</table>

NOTE: The numbers in parentheses are the corresponding standard errors (of RPE).
Simulation Comparison

- Lasso is the best when $\sigma$ is largest.
- At medium and large values of $\sigma$ Adaptive Lasso outperforms both SCAD and the garotte.
- Overall, Adaptive Lasso seems to do the best.
- No one method dominates the others, and they all seem to have specific strengths.
Summary

- $L_1$ penalization methods prove to be competitive with other methods in terms of prediction accuracy, and have the added benefit of variable selection, and thus contributing interpretable models.

- Furthermore, $L_1$ penalization have the advantage over subset selection procedures in that they are continuous, rather than discrete procedures.

- The lasso procedure does not *in general* have the oracle properties, but must satisfy a non-trivial necessary condition in order to attain these properties.

- Adaptive Lasso is an oracle procedure.

- When comparing the two $L_1$ penalization methods with the garotte and SCAD, still no one method can definitively be declared the best method universally.